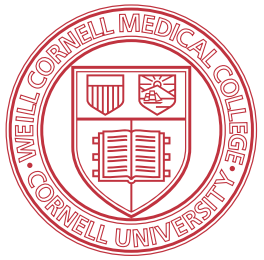
Differential gene expression analysis using RNA-seq

Applied Bioinformatics Core, September/October 2018



Friederike Dündar with Luce Skrabanek & Paul Zumbo

Day 2: Aligning reads

1. Experimental Design
2. Reference genome & transcript annotation
3. Alignment
   - STAR
   - BAM/SAM files
4. QC

EXPERIMENTAL DESIGN

How to avoid spurious signals and drowning in noise

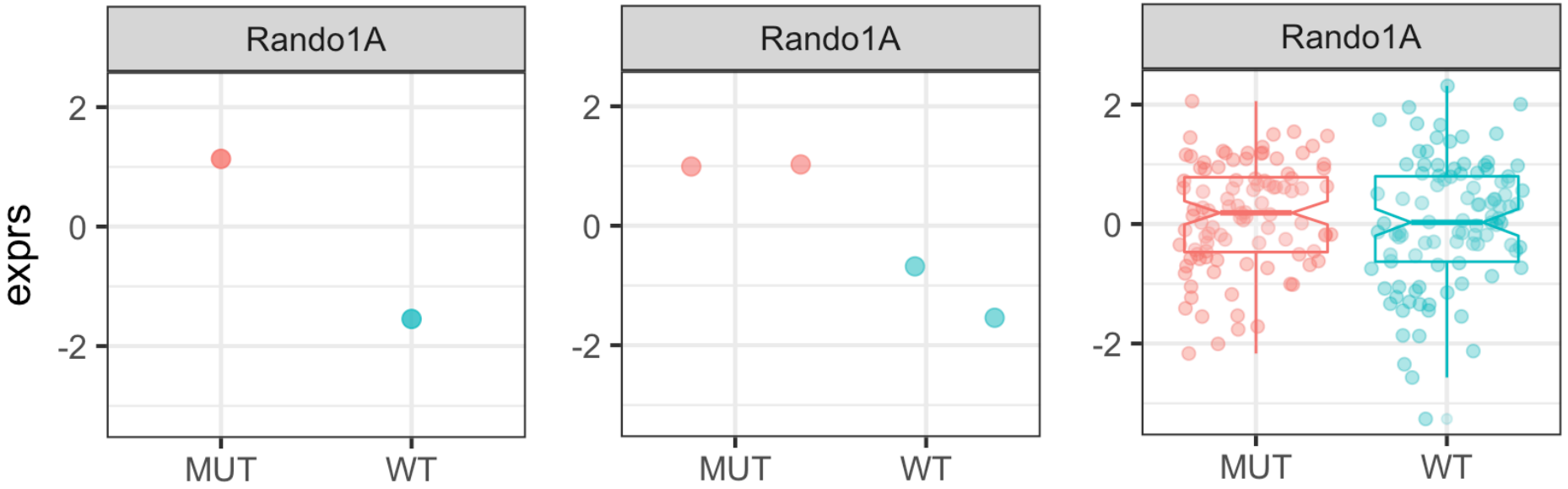
Why do we need replicates?

Goal: Identify differences in expression for every gene.

…and "differences" should preferably be due to our experiment, not noise!

"Samples are our windows to the population, and their statistics are used to estimate those of the population."
Martin Krzywinski & Naomi Altman

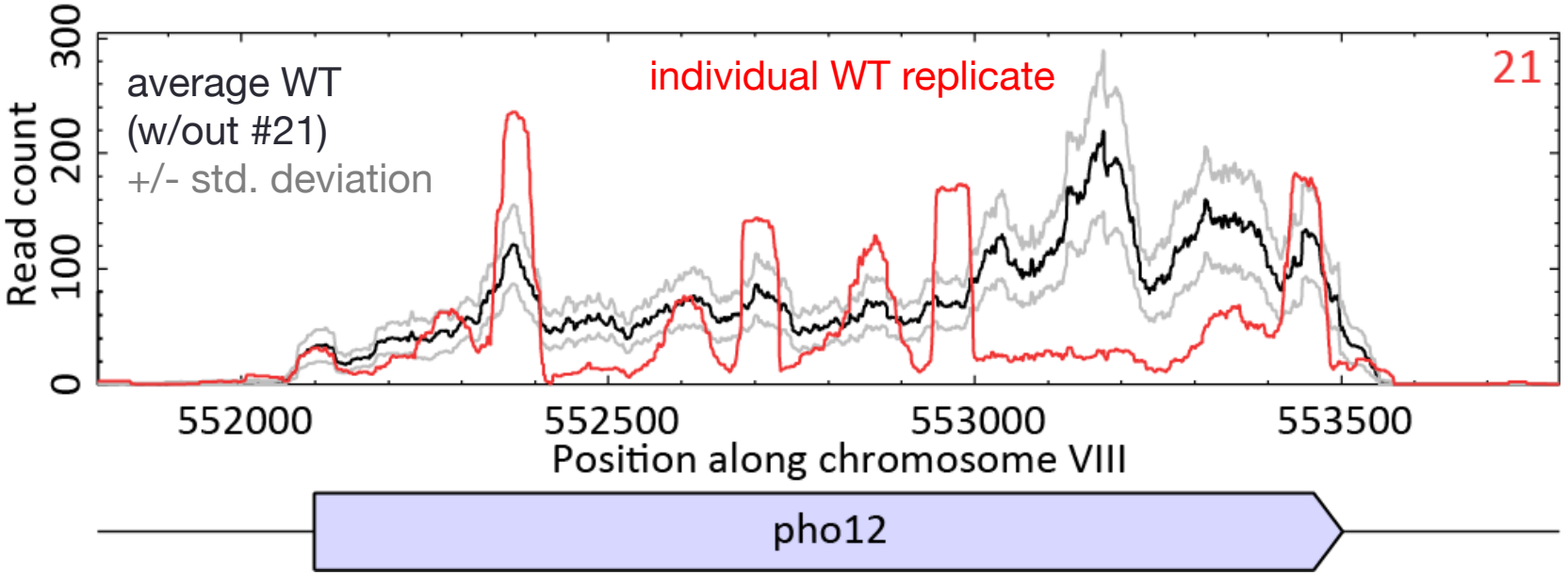
```
testdat <- data.frame(exprs = rnorm(200),
                      condition = c("WT","MUT"),
                      gene_name = "Rando1A")
```



doi:10.1038/nmeth.2613

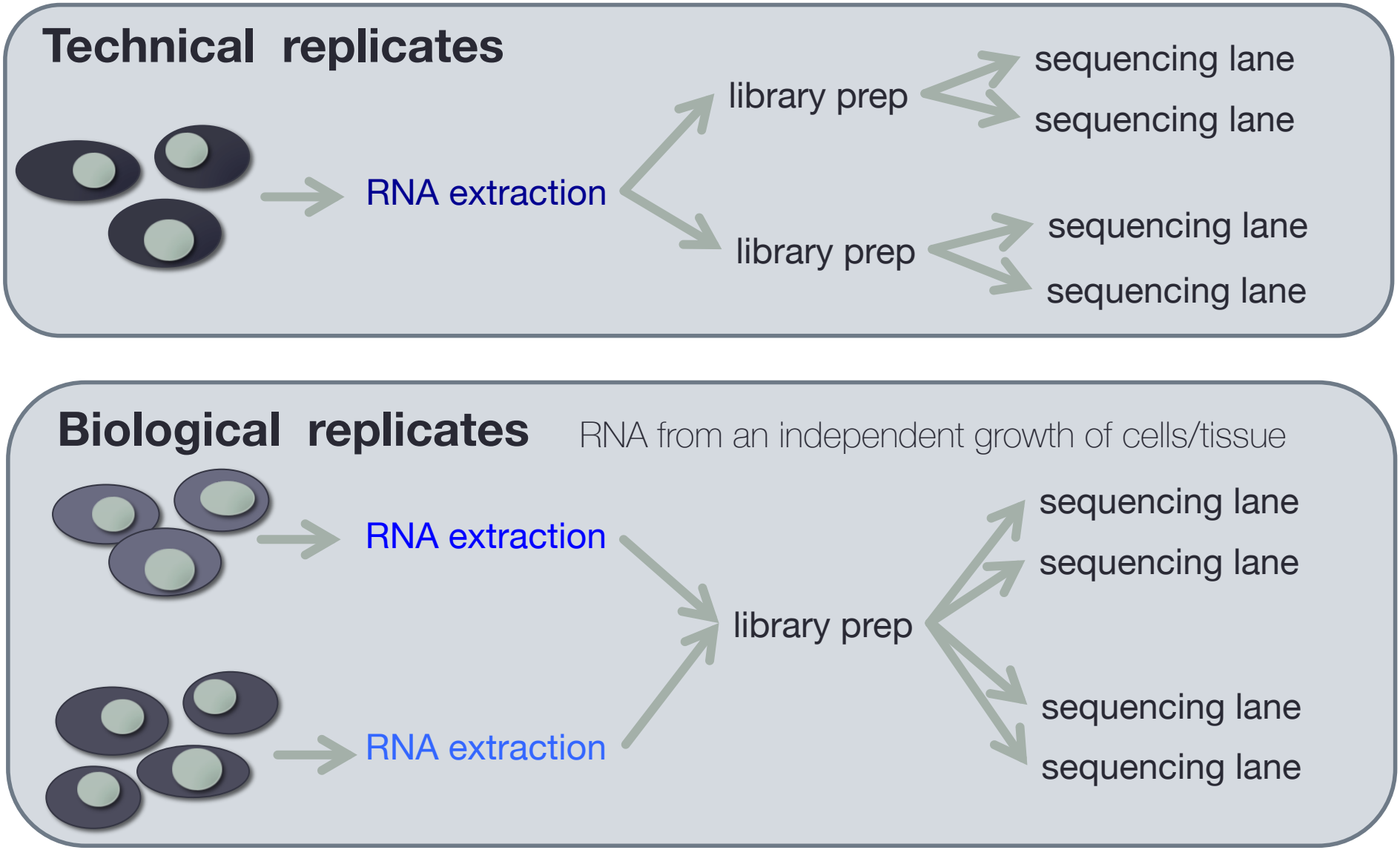
Invest in replicates!

- recommended: **6 biological replicates per condition** for DGE of strongly changing genes (logFC >= 2) [based on insights from the fairly simple yeast transcriptome]



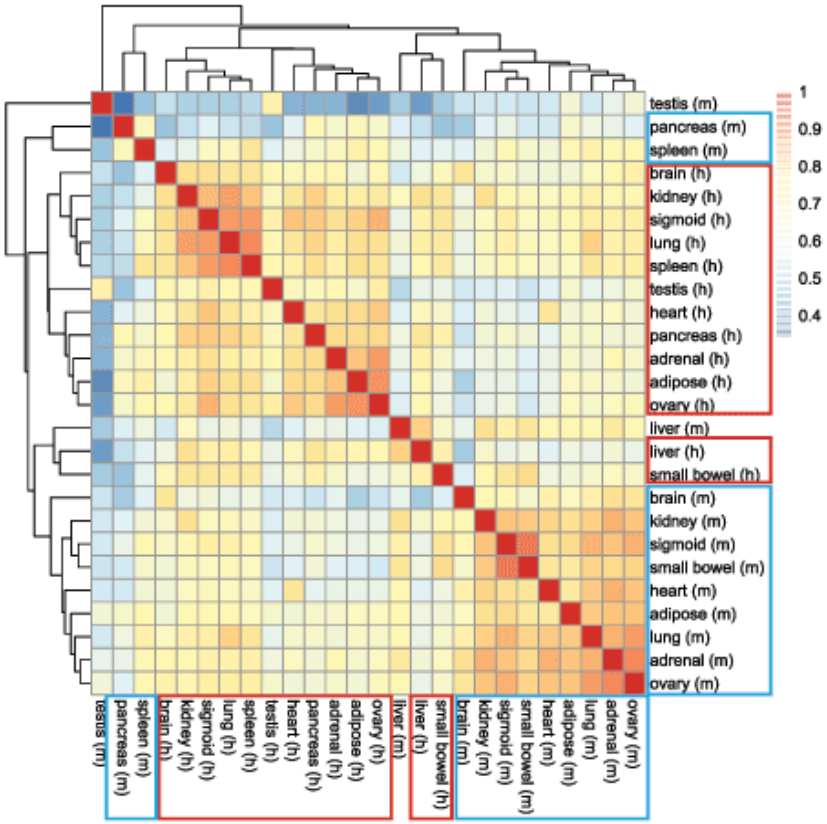
The most effective way to improve detection of differential expression in low expression genes is to add more biological replicates, rather than adding more reads (see Rapaport et al., 2013)

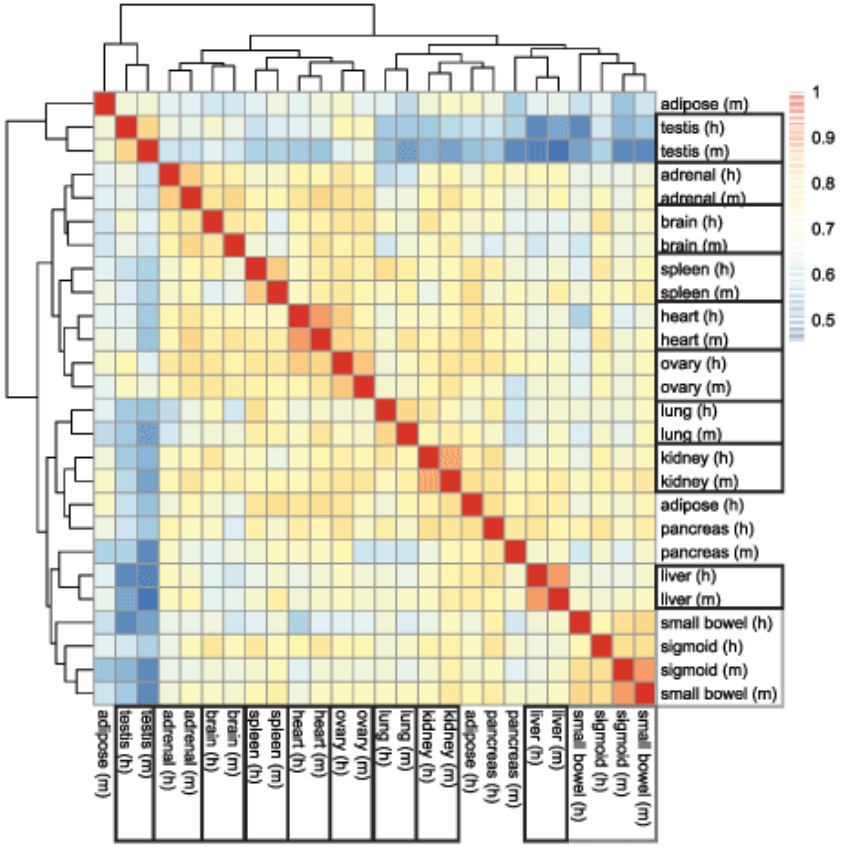
Replicate types



also see course notes and Blainey et al. (2014) Nature Methods, 1(9) 879–880.

Batch effects can happen everywhere





"Overall, our results indicate that there is ***considerable RNA expression diversity between humans and mice****, well beyond what was described previously, likely reflecting the fundamental physiological differences between these two organisms. "*

*"****Once we accounted for the batch effect*** *(…), the comparative gene expression data no longer clustered by species, and instead, we observed* ***a clear tendency for clustering by tissue****."*

Lin, Lin, and Snyder (2014). PNAS 111:48

Gilad & Mizrahi-Man (2015). F1000Research 4:121

ENCODE's* study design was not optimal

Most human samples were sequenced separately from the mouse samples:

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

Many tissues were not sex-matched

Tissue	Human	Mouse
adipose	FEMALE	MALE
adrenal	MALE	FEMALE
brain	FEMALE	MALE
heart	FEMALE	FEMALE
kidney	MALE	FEMALE
liver	MALE	FEMALE
lung	FEMALE	FEMALE
ovary	FEMALE	FEMALE
pancreas	FEMALE	FEMALE
sigmoid colo	MALE	FEMALE
small bowel	FEMALE	FEMALE
spleen	FEMALE	MALE
testis	MALE	MALE

not all variables can be controlled for
human data: deceased organ donors
mouse data: 10-week-old littermates

and that's ok, but you got to be mindful of these limitations when making bold claims

A very good read (including the reviews and comments) that discusses many scientific as well as ethical issues: https://f1000research.com/articles/4-121/v1

* not just ENCODE: see e.g. Leek et al. (2010) Nat Rev Gen 11(10) 733-739 or Jaffe & Irizarry (2014) Genome Biol 15(R31) 1–9

Avoiding bias

Completely randomized design

STRESS	A	B	A	A	B	A	B	A	A	B	B	B
DIET	1	2	1	2	2	1	1	2	2	1	2	1

Restricted randomized design

GENOTYPE	A	A	A	A	A	A	B	B	B	B	B	B
DIET	1	2	1	2	2	1	1	2	1	1	2	2

Blocked & randomized design

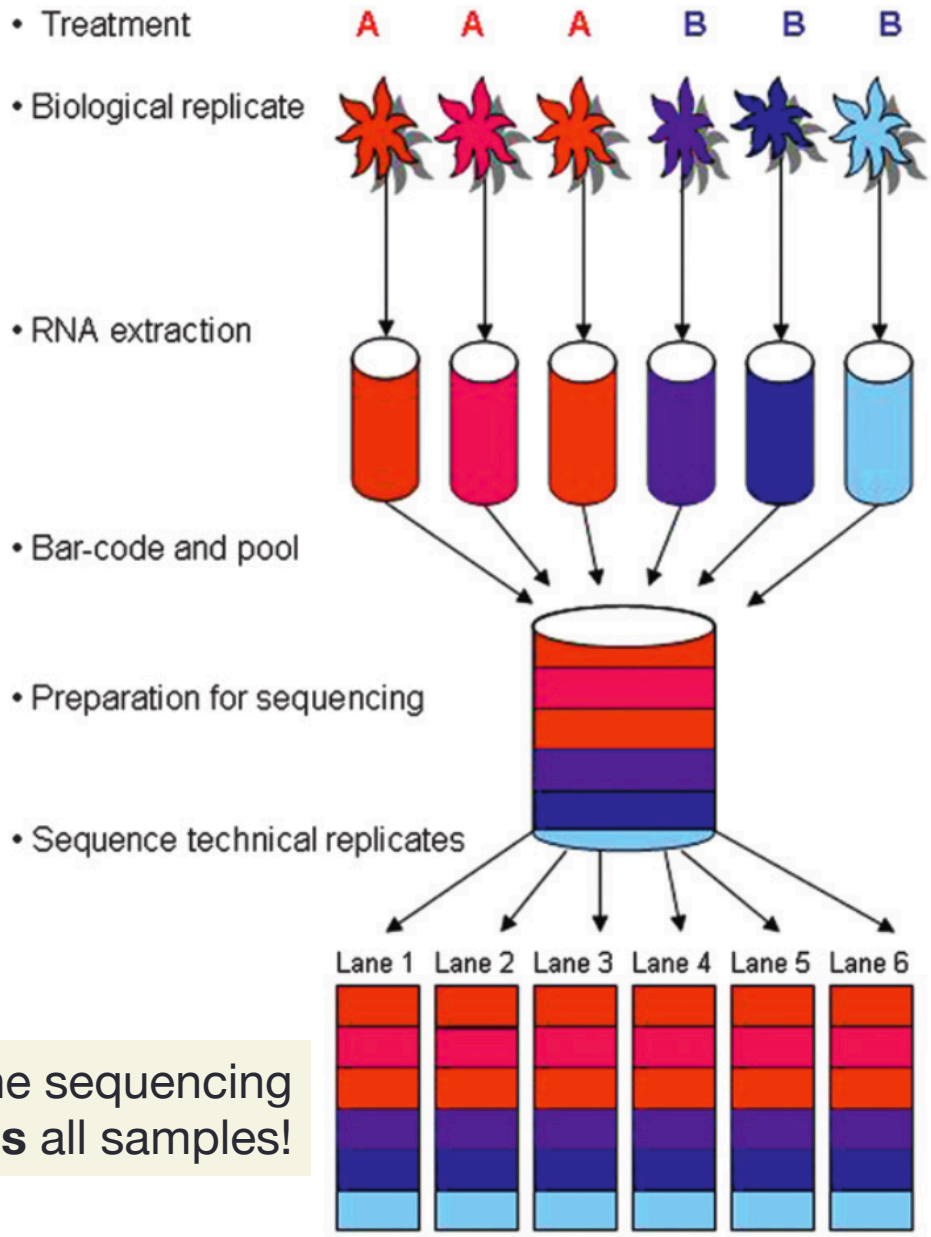
GENOTYPE	A	A	B	B	A	A	B	B	A	A	B	B
DIET	1	2	1	2	1	2	1	2	1	2	1	2
WEIGHT	•	•	•	•	●	●	●	●	●	●	●	●

Block what you can,
randomize what you cannot.

*What factors are of **interest**? Which ones might introduce noise?*
Which nuisance factors do you absolutely need to account for?

Typical RNA-seq set-up

- keep the **technical nuisance** factors (harvest date, RNA extraction kit, sequencing date…) to a **minimum**
- cover only as much of the **biological variation as needed** (but keep possible limitations for the final conclusions in mind)



Make sure the sequencing core **multiplexes** all samples!

How deep is deep enough?

for DGE (logFC~ 2) in mammals:
20 – 50 mio SR, 75 bp

Goals that require **more**, **longer,** and possibly **paired-end** reads:

- quantification of **lowly expressed** genes
- identification of genes with **small changes** between conditions
- investigation of **alternative splicing**/isoform quantification
- identification of **novel transcripts**, chimeric transcripts
- *de novo* **transcriptome assembly**

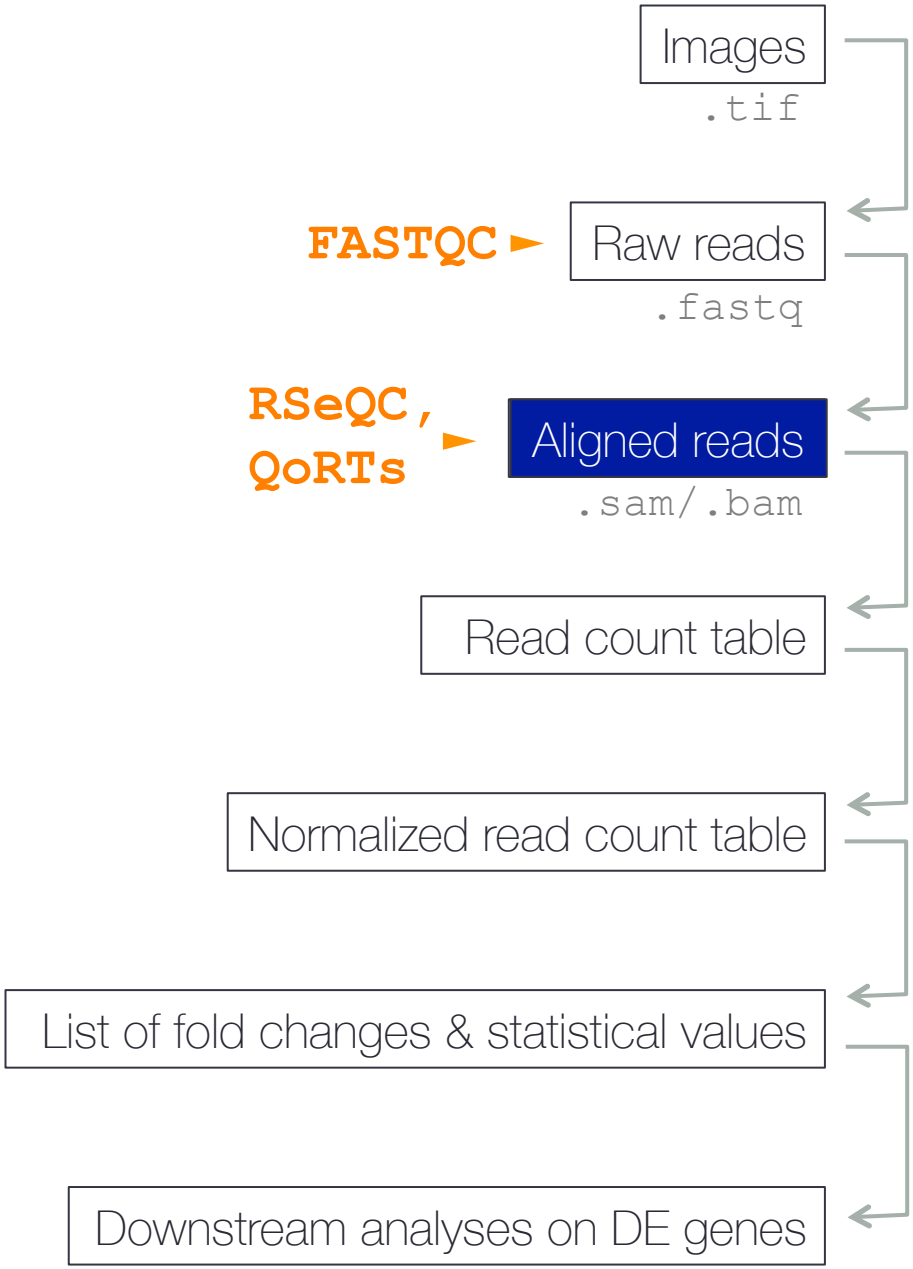
Remember: The addition of replicate samples provides substantially greater detection power of DE than increased sequence depth. (Rapaport et al., 2013)

Summary I

- RNA-seq analysis is not a completely solved issue – but **DE analysis on a gene level** is decently mature and the field seems to gravitate towards some sort of standard
- no analysis tool can enforce (or replace!) common sense and knowledge about the biology behind the experiment
- crap in, crap out
- more replicates are often better investments than more reads
- `FastQC` and `multiQC` are great tools that you should use

ALIGNMENT

Finding out where the reads came from



Base calling & demultiplexing

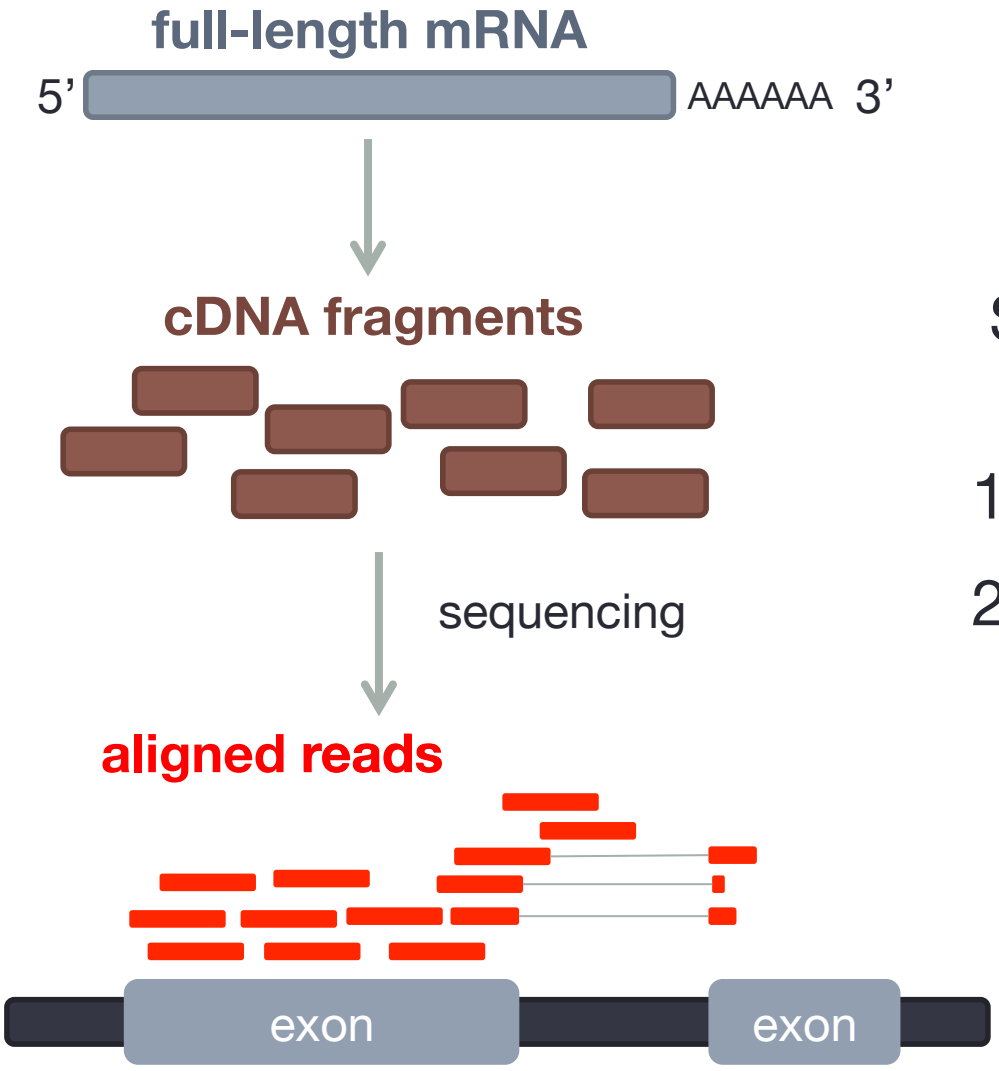
`Bustard/RTA/OLB, CASAVA`

Mapping

STAR

1. Theoretical background
2. Annotation files
3. Files with aligned reads: `SAM/BAM`
4. Looking at aligned reads

Aligning short RNA-seq reads

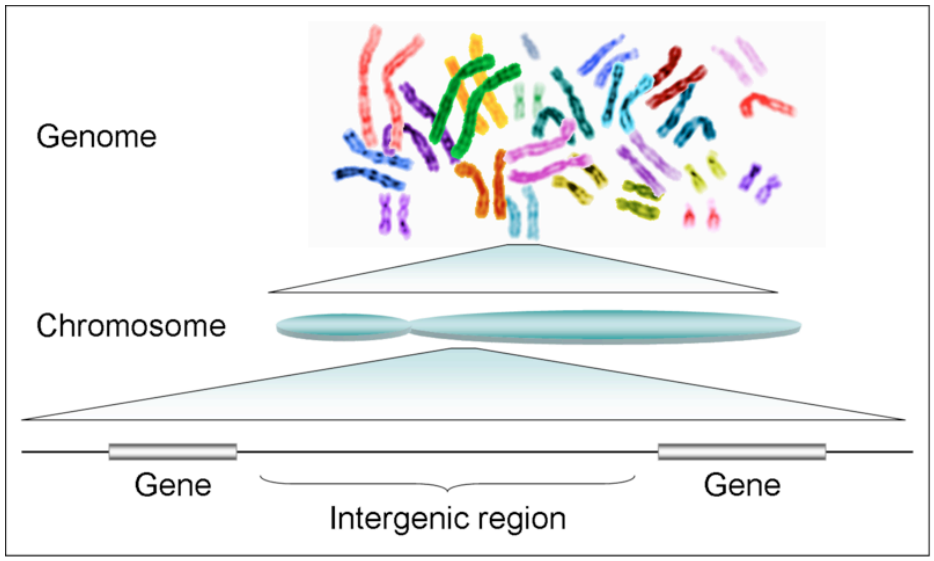


Spliced alignment tools usually need:

1) reference genome

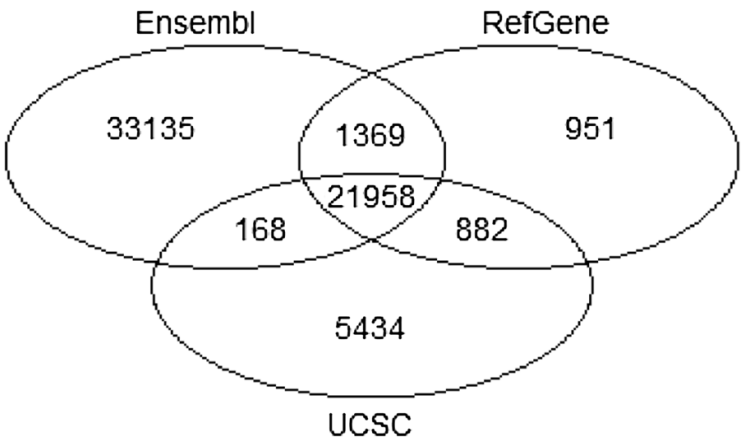
2) annotation

Annotation & their data bases



https://commons.wikimedia.org/wiki/File:Human_genome_to_genes.png

- Annotation is **dynamic**! (sequence, coordinates, types of elements)
- Automated vs. manual curation ("evidence-based")



RefSeq ncbi.nlm.nih.gov/refseq

UCSC Known Genes genome.ucsc.edu

Ensembl/Gencode gencodegenes.org

1/3 protein-coding genes
> 17,000 non-coding RNAs
> 15,000 pseudogenes

Zhao & Zhang (2015) BMC Genomics. doi:10.1186/s12864-015-1308-8

Which annotation should one use?

*"More sensitive annotations, such as **Ensembl** (…) **should be preferred** over more specific annotations, such as RefSeq (…) if the aim is to obtain accurate expression estimates."*

Janes et al. (Briefings in Bioinformatics, 2015). doi: 10.1093/bib/bbv007

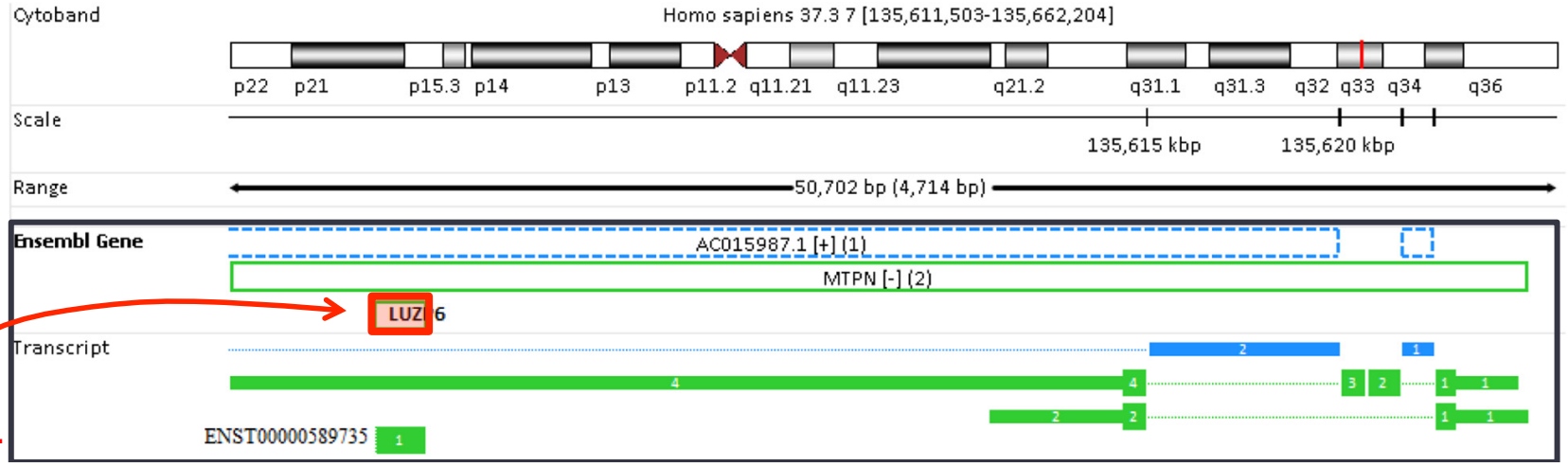
*"We observe that **RefSeq Genes produces the most accurate fold-change measures** with respect to a ground truth of RT-qPCR gene expression estimates. "*

Wu et al. (BMC Bioinfo, 2013). doi: 10.1186/1471-2105-14-S11-S8

*"In practice, there is **no simple answer to this question**, and it depends on the purpose of the analysis. (…) When choosing an annotation database, researchers should keep in mind that **no database is perfect** and **some gene annotations might be inaccurate or entirely wrong**."*

Zhao & Zhang (BMC Genomics, 2015). doi:10.1186/s12864-015-1308-8

Gene models can differ dramatically



vs.

1

2

1 **ENSEMBL**
0 unambiguous reads for LUZP6;
x number of reads for MTPN

vs.

2 **RefSeq**
0 reads for either
or
the same number of reads
for both genes

pay attention to the **source** as well as to the version of the genome/annotation **build**!

Storing annotation information

- representing genome coordinates + description/name
- various formats (all are plain text files): GFF2, GFF3, GTF, BED, SAF…

see the course notes for details

GTF ("GFF2.5")

1. reference coordinate
2. source
3. annotation type
4. start position
5. end position
6. score
7. strand
8. frame/phase
9. attributes: <TYPE VALUE>; <TYPE VALUE>; <TYPE VALUE>

GFF2

```
# GFF-version 2
IV      curated exon    5506900 5506996 . + .   Transcript B0273.1
IV      curated exon    5506026 5506382 . + .   Transcript B0273.1
IV      curated exon    5506558 5506660 . + .   Transcript B0273.1
IV      curated exon    5506738 5506852 . + .   Transcript B0273.1

# GFF-version 3
ctg123  .  exon  1300  1500  .  +  .  ID=exon00001
ctg123  .  exon  1050  1500  .  +  .  ID=exon00002
ctg123  .  exon  3000  3902  .  +  .  ID=exon00003
ctg123  .  exon  5000  5500  .  +  .  ID=exon00004
ctg123  .  exon  7000  9000  .  +  .  ID=exon00005
```

GFF3

GTF

```
# example for the 9th field of a GTF file
    gene_id "Em:U62.C22.6"; transcript_id "Em:U62.C22.6.mRNA"; exon_number 1
```

0 vs. 1 based conventions

one-based, fully-closed

ATG location: 7 - 9 or [7,9]
Cut site: 11^12 or (11,12)
Interval length = stop - start + 1

GFF format

zero-based, half open

ATG location: 6 - 9 or [6,9)
Cut site: 11-11 or [11,11)
Interval length = stop - start

BED format

http://alternateallele.blogspot.com/2012/03/genome-coordinate-cheat-sheet.html

Different philosophies of transcript quantification

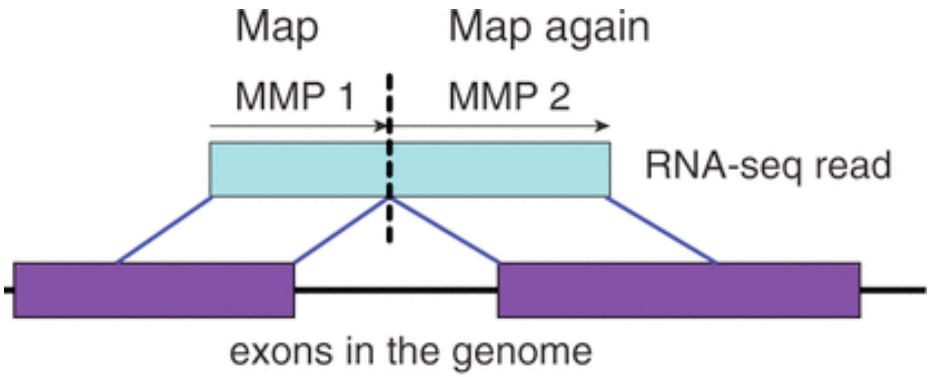
- (splice-aware) **alignment** followed by **counting** of reads overlapping with a given gene
  - STAR, TopHat followed by HTSeq-Counts, featureCounts

- (splice-aware) **alignment** followed by identification of the minimal number of **transcripts** that are supported by reads aligning to a given locus
  - Cufflinks

please do **not** use the Tuxedo suite!

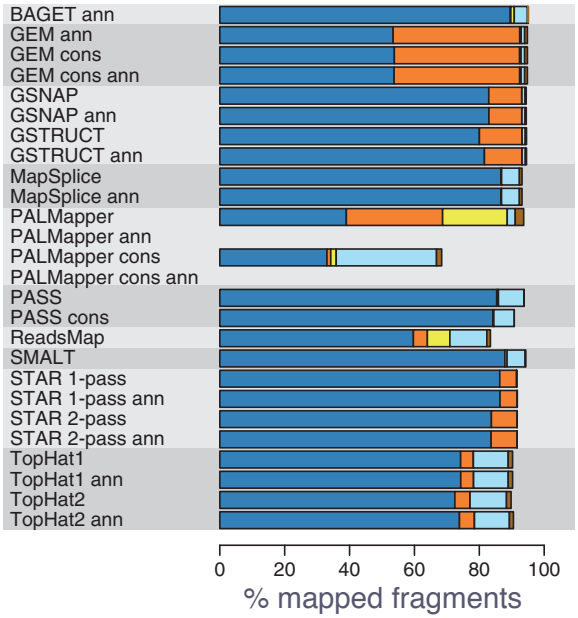
- **quantify** the number of reads that are consistent with a given transcript (the exact location within the transcript is ignored)
  - kallisto, salmon

Spliced Transcriptome Alignment to Reference (STAR)



- accurate & sensitive
- very fast
- memory intensive! (use it on the server!)

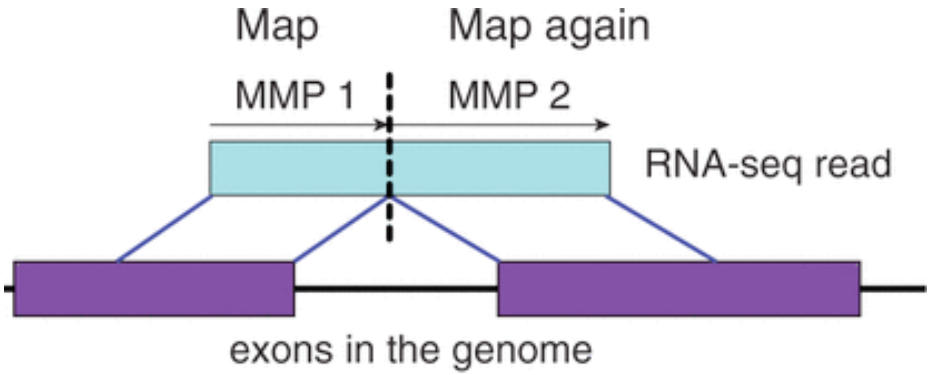
Spliced alignment programs



Engström et al. (2013) Nature Methods, 10(12), 1185–1191. doi:10.1038/nmeth.2722

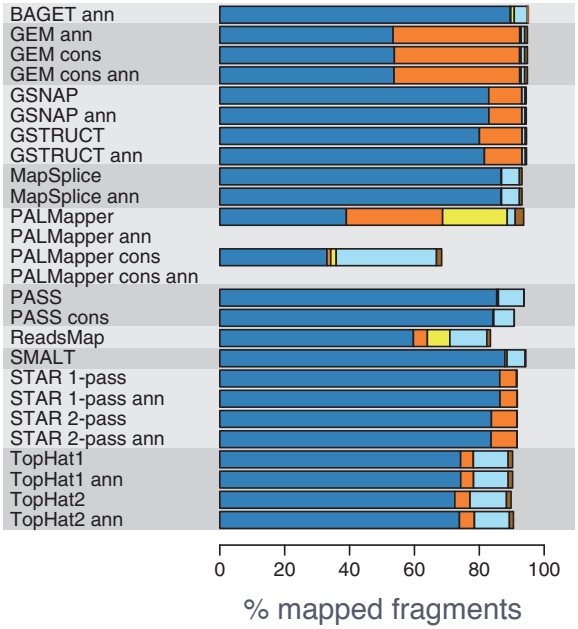
- MMP = maximal mappable prefix (aka *maximum matching portion*)
- reads are split when a continuous alignment is not possible
- the remaining unmappable portion is then aligned again
- finally, aligned portions of the original full-length reads are stitched together

STAR spliced alignment



- accurate & sensitive
- very fast
- memory intensive!

Spliced alignment programs



Engström et al. (2013) Nature Methods, 10(12), 1185–1191. doi:10.1038/nmeth.2722

STAR has myriad options! Tune them to meet your needs

Current Protocols in Bioinformatics (Sept 2015)
DOI: 10.1002/0471250953.bi1114s51
and
STARmanual.pdf

2 main STAR modules

1. generate **genome index**

```
--runMode genomeGenerate
--genomeFastaFiles sacCer3.fa
--sjdbGTFfile sacCer3.gtf
```

needs to be done just 1x per transcriptome!

2. **align**

   2.1. align to *reference* & identify novel splice junctions

   2.2 re-run alignment including the novel splice junctions

```
$runSTAR –genomeDir STARindex/ \
         --readFilesIn $FASTQ_FILES \
         --readFilesCommand zcat \
```

```
--twopassMode
```

must be done for every sample

Let's align the reads for WT_1!

Storing aligned reads: SAM/BAM

@HD VN:

@SQ SN: LN:

@RG ID: SM:

@PG ID:

@CO

(theoretically) optional
HEADER SECTION
general information about the file

1	2	3	4	5	6	7	8	9	10	11	>11
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

Paired read?
Unmapped?
Mapped to rev. strand?
1st in pair?
2nd in pair?
Failed QC?
…

M (mis)match
I insertion
D deletion
N skipped
S soft clipped
H hard clipped
P padding

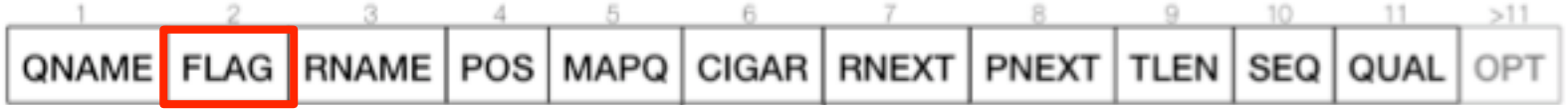
<TAG>:<TYPE>:<VALUE>

AS	A
BC	i
NH	f
NM	z
…	H

ALIGNMENT
SECTION
1 line per locus

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

Storing aligned reads: SAM/BAM



2nd field: binary FLAG

Binary (Decimal)	Hex	Description
00000000001 (1)	0x1	Is the read paired?
00000000010 (2)	0x2	Are both reads in a pair mapped "properly" (i.e., in the correct orientation with respect to one another)?
00000000100 (4)	0x4	Is the read itself unmapped?
00000001000 (8)	0x8	Is the mate read unmapped?
00000010000 (16)	0x10	Has the read been mapped to the reverse strand?
00000100000 (32)	0x20	Has the mate read been mapped to the reverse strand?
00001000000 (64)	0x40	Is the read the first read in a pair?
00010000000 (128)	0x80	Is the read the second read in a pair?
00100000000 (256)	0x100	Is the alignment not primary? (A read with split matches may have multiple primary alignment records.)
01000000000 (512)	0x200	Does the read fail platform/vendor quality checks?
10000000000 (1024)	0x400	Is the read a PCR or optical duplicate?

most common FLAGS for SR: 0; 4; 16

`https://broadinstitute.github.io/picard/explain-flags.html`

Storing aligned reads: SAM/BAM

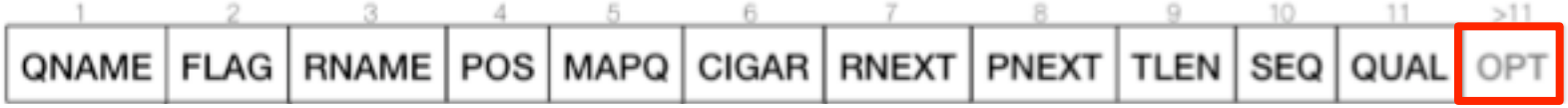
6th field: CIGAR string – which hoops did the aligner have to jump through to align the read to the genome locus that it thought was the best fit?

M	alignment (match or **mis**match!!)
I (N)	insertion (large insertions)
D	deletion
S/H	clipping

spliced out introns = sequences are missing in the read, i.e., they need to be inserted in order to align the read to the genome

Reference sequence with aligned reads	CIGAR string	Explanation
C T G C A T G T T A G A T A A * * G A T A G C T G T G C T A		
reads: A A G G A T A * C T G	1M2I4M1D3M	Insertion & Deletion
reads: G A T A A * G G A T A	5M1P1I4M	Padding & Insertion
reads: T G T T A ..... T G C T A	5M15N5M	Spliced read
reads: a a a C A T G T T A G	3S8M	Soft clipping
reads: A A A C A T G T T A G	3H8M	Hard clipping

Storing aligned reads: SAM/BAM



after 11th field: OPTIONAL information

<TAG>:<TYPE>:<VALUE>
tags are not standardized!

`AS:i`	Alignment score
`BC:Z`	Barcode sequence
`HI:i`	Query is i-th hit stored in the file
`NH:i`	Number of reported alignments for the query sequence
`NM:i`	Edit distance of the query to the reference
`MD:Z`	String that contains the exact positions of mismatches (should complement the `CIGAR` string)
`RG:Z`	Read group (should match the entry after `ID` if `@RG` is present in the header.

`NH HI NM MD` have standard meaning as defined in the SAM format specifications.

`AS` id the local alignment score (paired for paired-edn reads).

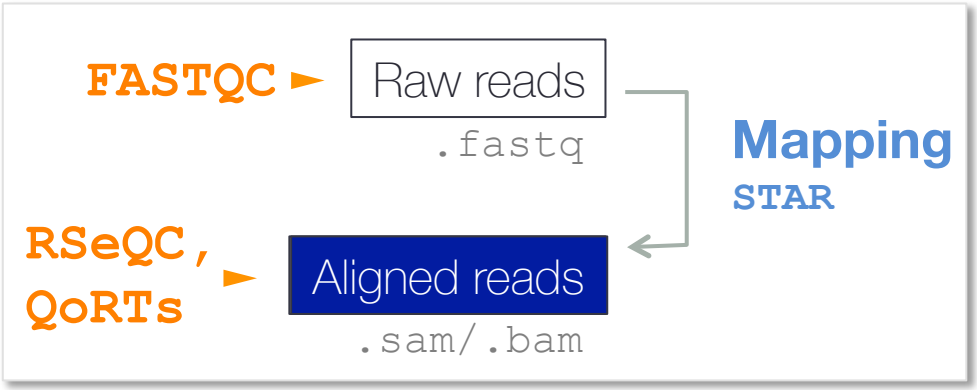
`nM` is the number of mismatches per (paired) alignment, not to be confused with `NM`, which is the number of mismatches in each mate.

`jM:B:c,M1,M2,...` intron motifs for all junctions (i.e. N in CIGAR): 0: non-canonical; 1: GT/AG, 2: CT/AC, 3: GC/AG, 4: CT/GC, 5: AT/AC, 6: GT/AT. If splice junctions database is used, and a junction is annotated, 20 is added to its motif value.

`jI:B:I,Start1,End1,Start2,End2,...` Start and End of introns for all junctions (1-based).

`jM jI` attributes require samtools 0.1.18 or later, and were reported to be incompatible with some downstream tools such as Cufflinks.

Recap



`day02_recap.pdf`

- We **downloaded fastq.gz** files from the SRA via ENA using wget
- We did QC of the raw reads using **FastQC** (1x per sample) and summarized it using **MultiQC**
- We **aligned** the raw reads using **STAR** (the genome index that is necessary was provided by us)
- We will do **additional QC** on those BAM files

Basic QC of aligned reads

How many reads were aligned? What were reasons for lack of alignment?
Do you have enough paired mates (for PE sequencing)?

- aligner output (e.g., Log.final.out)
- `samtools flagstat`
- RSeQC's `bam_stat`
- QoRTs

visual inspection!

(almost) all of these can be summarized using MultiQC!
→ Section 3.4.1 of the course notes

Integrative Genomics Viewer

http://software.broadinstitute.org/software/igv/download

Integrative Genomics Viewer (IGV) (Version 2.3)

Install IGV

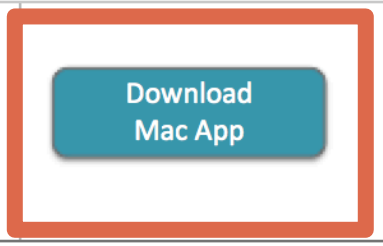
Options for installing and running IGV:

1. (Mac only) Download and run the Mac application; or
2. (Windows) Download and run the self-extracting archive; or
3. (All systems) Use the Java Web Start buttons (Mac users: see below for limitations); or
4. (All systems) Download the binary distribution and run IGV from the command line.

Note: IGV 2.3.x requires Java 7. Users with Java 6 (JRE 1.6) should first try to upgrade Java to the latest version. If this is not possible you will need to run a 2.2.x version available in the archive.

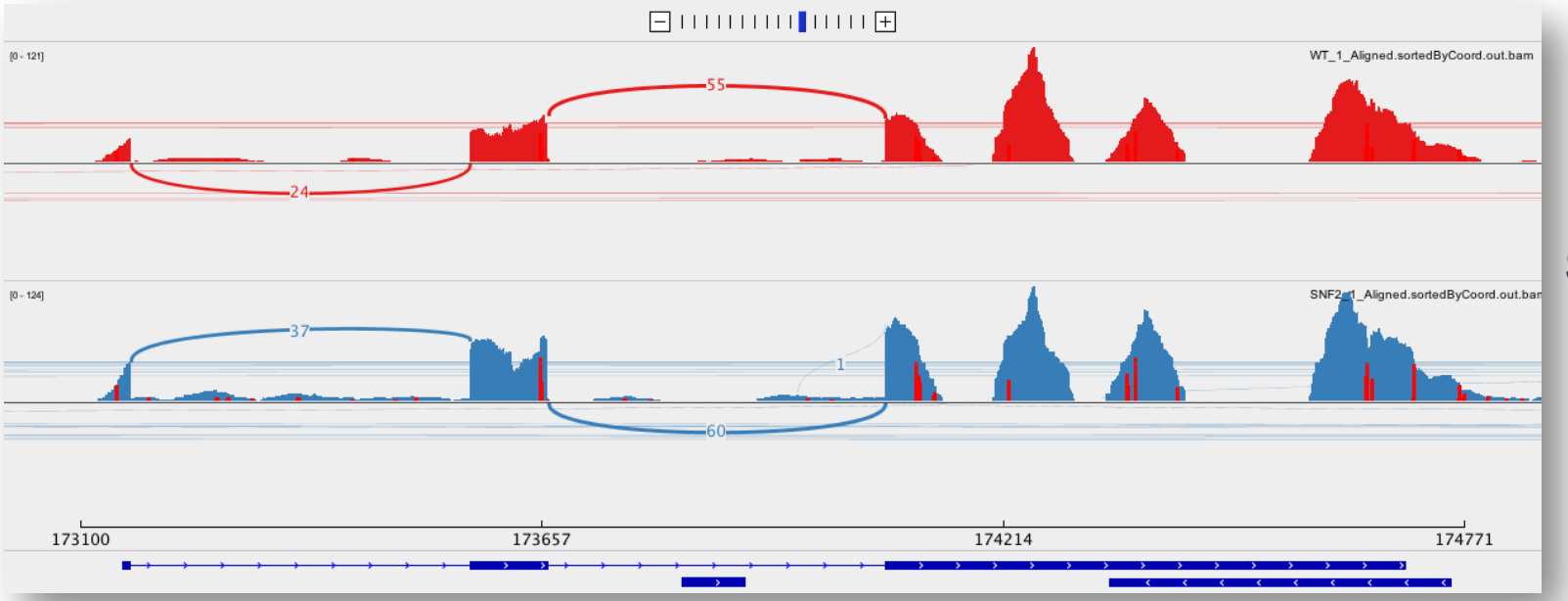
Mac

Download and unzip the Mac App archive, then double-click the IGV application to run it. The application can be moved to the "Applications" folder, or anywhere else.



Integrative Genomics Viewer

- load BAM file(s) from URL ("File" -> "Open URL…"):
  http://chagall.med.cornell.edu/RNASEQcourse/
  http://www.trii.org/courses/rnaseq.html
- take a snapshot of the reads around gene *YPL198W*



starting with IGV 2.3, **Sashimi plots** can easily be created

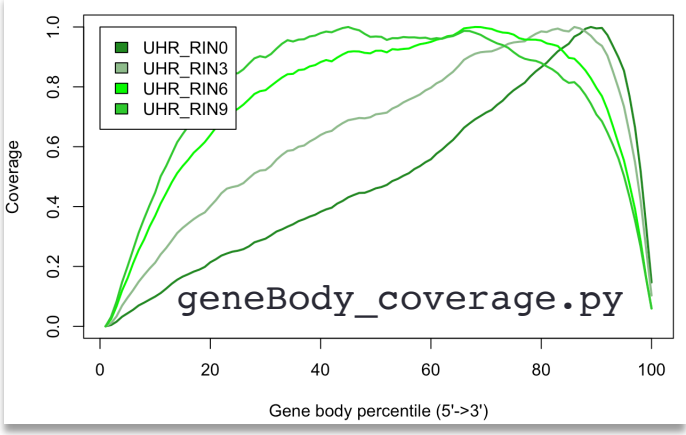
http://software.broadinstitute.org/software/igv/Sashimi

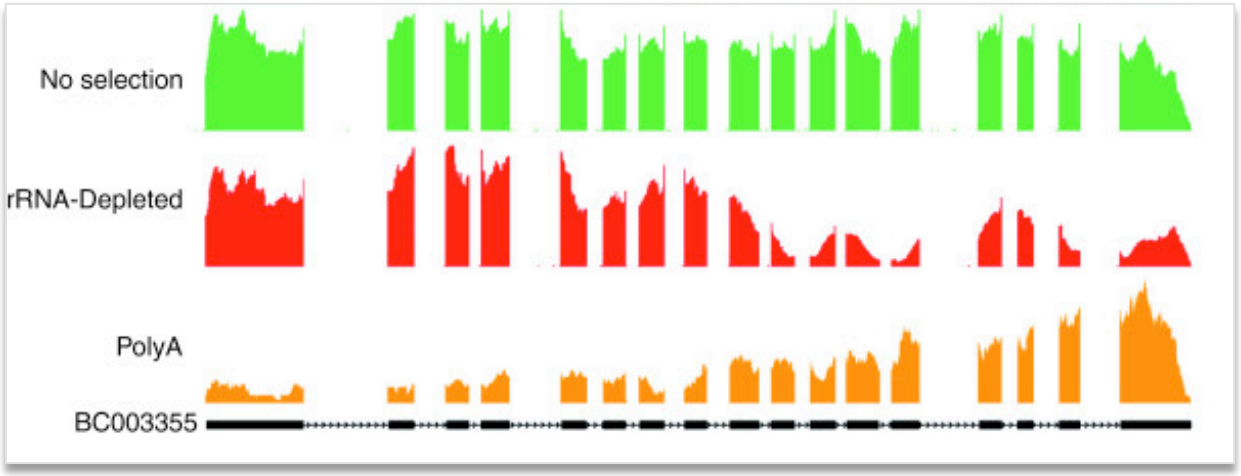
many options!

Katz (2015). Bioinformatics. doi:10.1093/bioinformatics/btv034

Identifying RNA-seq-specific biases

- lack of **gene diversity**: dominance of rRNAs, tRNAs (and other highly abundant transcripts)
- **read distribution**
  - high intron coverage ~ incomplete poly(A) enrichment step or many immature transcripts
  - many intergenic reads: possibly gDNA contamination
- **gene coverage: 3' bias** → RNA degradation, poly(A) enrichment protocol





popular QC tools:
RSeQC, QoRTs

(see course notes)

Lahens et al. (2014) Genome Biology 15:R86

removing rRNAs

Can be done at virtually every step of the analysis. Choose the version that makes most sense to you.

- **sortMeRNA**: http://bioinfo.lifl.fr/RNA/sortmerna/
  - input: reads in fastq file + rRNA sequences
  - will extract those reads that do not match to the rRNA sequences
  - https://www.ncbi.nlm.nih.gov/nuccore/U13369 (human rRNA), https://www.ncbi.nlm.nih.gov/nuccore/BK000964 (mouse)

raw reads filtering

- make a **"genome" index for rRNAs only** (and perhaps tRNAs), then align your reads and only use those that do not map for the next round of alignment

alignment-based filtering

- do your alignment and counting as is, simply **ignore the rRNA genes** in your subsequent downstream analysis

ignoring information about some RNA classes

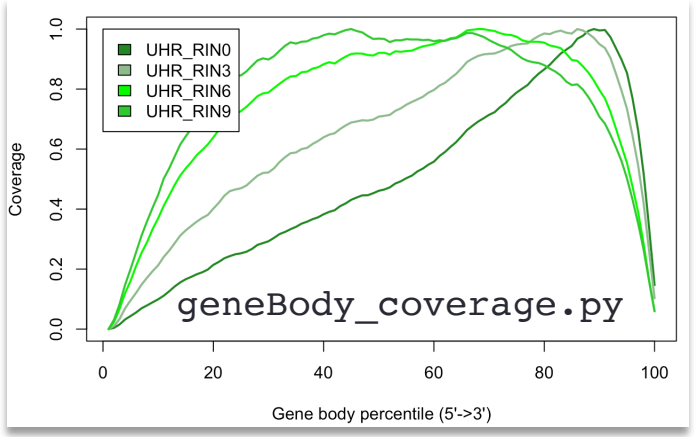
QC recap

- **raw reads** QC
  - adapter/primer/other contaminating and over-represented sequences
  - sequencing quality
  - GC distributions
  - duplication levels

FastQC
(QoRTs)

aligner's log files
samtools flagstat
RSeQC
QoRTs
MultiQC
…

- **aligned reads** QC
  - % (uniquely) aligned reads
  - % exonic vs. intronic/intergenic
  - gene diversity
  - gene body coverage



Summary II

- aligning unspliced reads is not too difficult, but it still takes a long time (depending on the size of the genome)
- spliced reads are quite tricky, and identifying novel splice junctions is error-prone and far from being solved
- the file format for storing aligned reads (SAM/BAM) is fairly standardized, but the optional fields (and how alignment tools interpret some of the mandatory entries) leave lots of room for variability
- the file format(s) for storing genome annotation (e.g. genes, transcripts) tend to be even stricter defined and even less well followed (aka it's a mess!)
- historically, samtools are the most widely used tools when it comes to exploring and manipulating SAM/BAM files (although there are alternatives, e.g. bamtools)
- **QC of aligned read files is at least as important as QC of the raw reads, if not more so!**